ChinaXiv合作期刊

双向模式匹配在年鉴数据预处理平台中的应用

史礼婷[1] 张 骞[2] 钟永恒[1] 胡思思[1] 李贞贞[1]
[1](中国科学院武汉文献情报中心 武汉 430071)
[2](中国航天科工集团第九总体设计部 武汉 430040)

摘要:【目的】实现年鉴指标数据的结构化存储，完成年鉴数据的更新录入。【应用背景】年鉴预处理平台是将年鉴数据统一整理、审核、上传的 C/S 工具平台，采用 VC++为主要编程语言，为年鉴数据库建设提供数据基础。【方法】双向模式匹配处理是在 WM 模式算法基础上进行改进，利用分词技术对录入指标进行信息元提取、采用存储过程实现模式集合的筛减、信息双向匹配保证匹配的准确高效。【结果】通过对实验数据录入的匹配结果进行分析，发现双向模式匹配有较高指标匹配率和正确率。【结论】双向匹配算法能满足年鉴录入的需求，提高了年鉴数据预处理工作的效率。

关键词: 双向模式匹配 年鉴数据 WM 算法
分类号: G350

chinaXiv:201711.02037v1

1 引 言

年鉴数据是一种全面、系统、准确地以记述上一时间段事物运动、发展状况为主要内容的事实资料汇集，包括综合性年鉴、专门性年鉴、统计性年鉴、地域性年鉴，对这些年鉴数据进行分析可以帮助人们了解事物现状和研究发展趋势，对于总结、统计和比较事物起到了参考作用，因此，广泛应用于各个研究领域，数据本身具有很大价值[1]。

现今，各种年鉴数据并没有统一格式规范，大多以数据表格的形式进行非结构化存储，再加上，国家和地域性数据在存储格式上存在较大差异，这些都不利于数据查询、处理以及分析。目前，国内外并没有公开可用的年鉴数据录入平台，若是单靠人力进行数据整理和录入，需要的时间和人力的成本是相当大的。为了保障数据的真实准确，充分利用现有年鉴资源，笔者建立了以海量年鉴数据为处理对象的年鉴数据预处理平台，利用计算机对格式多样的指标数据进行统一录入，将非结构化的年鉴数据批量存储到数据库中，形成结构化数据，规范了操作流程，从源头上保障了数据质量，为产业技术情报工作者提供了很好的数据分析基础，具有重要的现实意义。

平台中的自动匹配模块，可以将不同年份的表格数据之间，描述相同指标的数据合并，进行归一化存储，方便了指标数据的查询，解决了数据的更新问题，是年鉴数据自动录入能否成功实现的关键步骤。本文对指标数据自动匹配进行探索性研究，根据指标数据的特点，设计并实现了指标数据双向匹配，为年鉴数据自动录入工作提供了良好的技术支持，提高了年鉴数据预处理的工作效率。

2 平台设计与算法分析

2.1 平台的功能设计

年鉴数据预处理平台的处理对象为年鉴表格文件，考虑到不同类型年鉴处理过程的相对独立性，系统采用 Client/Server 架构实现，以应用工具的形式提

通讯作者：史礼婷，ORCID：0000-0002-7618-3777，E-mail：shilt@mail.whlib.ac.cn。

ChinaXiv合作期刊

chinaXiv:201711.02037v1

供给数据录入人员使用。将计算和数据合理地分配在客户机和服务器两端，充分发挥客户端 PC 的处理能力，有效地降低网络通信量和服务器运算量[2]。

为了实现年鉴数据归一化存储，系统需要实现指标体系自动构建、年鉴文档数据自动识别、年鉴数据匹配更新录入、数据审核及数据上传模块，如图 1 所示。本文将重点阐述匹配更新模块的实现。

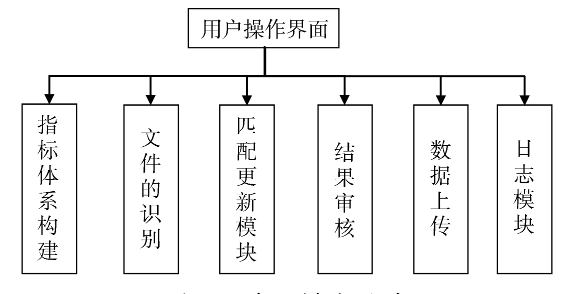


图 1　系统模块设计

2.2　模式匹配的研究现状

年鉴数据的匹配更新模块，是整个系统构建的核心和难点所在，其算法的优劣直接影响了数据的可持续性及录入工作的效率。多年指标数据往往存放于不同的指标表格文件中，在进行最新年份的指标数据录入的时候，需要通过描述指标的中文字符串的比对，找到其对应原始指标，从而进行最新年份的数据更新。如何能准确高效地实现中文指标字符的匹配成为问题关键。

中文字符串的匹配，又被称为字符串模式匹配，已有 40 多年的发展历史，是文本处理领域非常重要的一个研究课题。当前，由于网络信息的迅速增长，模式匹配技术已经广泛应用于各种领域，如过滤防火墙、网络搜索引擎、入侵检测系统、生物医学[3-4]等，并在多数操作系统和应用软件中也得到了实际应用。

BF(Brute Force)算法是最早出现的一种字符串模式匹配算法，是最基本最简单的匹配算法，又被称为蛮力算法。1977 年出现的 KMP(Knuth-Morris-Pratt)算法，是第一个时间复杂度达到 O(n)的算法[5]。另外一个著名的算法是 BM(Boyer-Moore)算法[6]，虽然最坏情况的时间复杂度是 O(mn)，但其在绝大多数场合的性能表现，比 KMP 算法还要出色。在假定字符等概率出现并且相互独立的条件下，该算法的平均时间复杂度下界是 $O(n\log_{|\sum m/m|})$，已经在 1979 年被 Yao 证明[7]。

在已有的多模式匹配算法中，较为经典的有 AC 算法、WM 算法。近年来，字符串匹配的研究领域也在不断发展，不断涌现出新的匹配算法。据不完全统计，2000 年之前公开发表的有近 40 种经典算法，2000 年之后陆续发表有 50 多种新算法[8]。

针对字符串匹配问题，学术界达成一个共识，往往思想越简单的算法，实际应用的性能越好。一部分学者对于模式匹配的研究存在理论脱离实际的现象，算法的复杂度不断提高，但是实际应用的效果却不是很理想，往往没有经典的 AC 算法、WM 算法更高效[9]。另一方面，随着网络信息的不断增长，需要处理的数据量变得越来越大，而人们对匹配速度的要求越来越高，这使得研究者面临着巨大挑战[4]。因此，在模式匹配的研究过程中，改进经典算法的同时，更要着眼于实际的应用效果，将理论与实际相结合，研究出适合大规模数据集进行模式匹配的方法，同时还要保证模式匹配效率。

2.3　WM 算法原理

多模式算法中，较为经典的有 AC 算法和 WM 算法，它们都需要对模式集进行预处理。AC 算法[10]需要维护一个状态机，所以在构建的时间和空间复杂度上，要比 WM 算法更消耗资源，而且，如果模式集动态可变，AC 算法动态调整自动机的成本要比 WM 算法高很多。WM 算法[11]利用跳跃思想使部分字符不需要进行匹配，同时采用 Hash 散列匹配的方法，提高了处理速度，在实际应用中取得了较高的效率。因此，在指标匹配的处理过程中，本文采用性能更好的 WM 算法。

WM 算法借鉴了 BM 单模式匹配算法中的坏字符跳转规则，而在多模式匹配问题中，由于模式串集合较大，覆盖的字符集也相应变大，导致坏字符出现几率变低，滑动窗口跳转距离减小，所以上述规则的效率明显下降。因此，WM 算法提出字符块这一概念，以几个连续的字符作为一个匹配单位，以此增加滑动窗口在每一次尝试匹配后的滑动距离，字符块长度 B 一般为 2 至 3 字节[11-12]。

WM 算法需要构建 SHIFT 表、HASH 表和 PREFIX 表，主要运用哈希表的思想提升性能。其中，SHIFT 表记录字符集中所有字符块在文本 T 中出现时滑动窗口的移动距离；HASH 表记录 SHIFT 项为 0 且后缀字符块哈希值相同的所有特征串位置；PREFIX 表则记录特征串集中前缀字符块哈希值相同的特征串位置。具

ChinaXiv合作期刊

chinaXiv:201711.02037v1

体的匹配过程如下[11-12]:

(1) p 指向文本 T 滑动窗口后缀。如果 $p > T_{end}$，退出；否则，计算 p 指向的字符块的 hash 值。

(2) 查 SHIFT 表，取 SHIFT[hash]值。如果等于 0，表示 p 指向的字符块与某个特征串后缀字符块相同，转步骤(3)；不为 0，则执行 p=p+SHIFT[hash]，转步骤(1)。

(3) 查 HASH 表，HASH[hash]指向可能与文本 T 当前滑动窗口内子串匹配的所有特征串。

(4) 计算文本 T 当前滑动窗口内子串前缀字符块的散列值 prefix。

(5) 对 HASH[hash]指向的每个特征串查 PREFIX 表，如果其 PREFIX[hash]与 prefix 值相同，则进行精确匹配，从特征串第一个字符进行完全匹配，报告结果。全部匹配过程结束，执行 p=p+1，转步骤(1) 。

3 双向模式匹配的实现

3.1 指标匹配的设计

指标数据更新模块，需要完成的历年数据的汇总。考虑到数据的严谨性，匹配结果的准确性是判断更新是否成功的首要标准，在此前提下，需要尽可能提高处理过程的效率。而将 WM 算法直接运用于指标匹配中是不可行的，主要是因为指标体系庞大，WM 预处理时间成本和空间成本大大增加，匹配性能急剧下降。针对指标匹配对准确性的要求和指标集合自身的特点，本文从以下方面对指标集合进行处理，使 WM 算法适用于指标的匹配，提高指标匹配的性能:

(1) 预先对指标模式集合进行一定程度的筛减，剔除掉大量无关指标集合。

(2) 对指标文本进行文本清洗，去除无关字符，提取表述指标的信息元集合来进行后续匹配。

(3) 在匹配环节加入反向匹配，进一步提高匹配的准确性及效率。

鉴于以上分析，本文双向模式匹配的流程设计如图 2 所示。

3.2 指标文本清洗及信息元的提取

双向匹配处理的第一步，就是指标文本清洗及信息元的提取，其结果直接影响后续中文字符匹配的准确程度。年鉴数据是以 Excel 文本文件进行存储的，需要对表格中表述指标的关键信息进行提取。年鉴数据

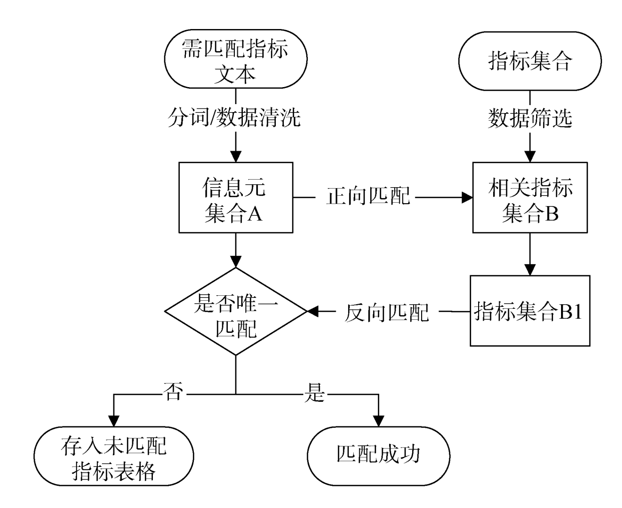


图 2 双向模式匹配处理的流程

表名、行列标题、单位这些指标名称关键字段中，常常会出现中文字符的错误，而这些错误符号或无用信息会直接影响指标的匹配。例如：固定资产投资 2014，这里 2014 属于年份信息，而不是指标名称，需要去除；又如：固定资产投资(，其中“(”则为错误符号，也是需要去除的部分。因此，在进行数据匹配之前，对于从 Excel 中提取的字符，要进行字符串的清洗，并提取描述指标的关键信息来进行后续的匹配，以确保字符匹配的准确性。

根据年鉴指标名称的特点，本文设计并实现了指标关键信息提取的操作流程，如图 3 所示，包括中文识别、中文分词、数据清洗、有用信息提取等步骤。

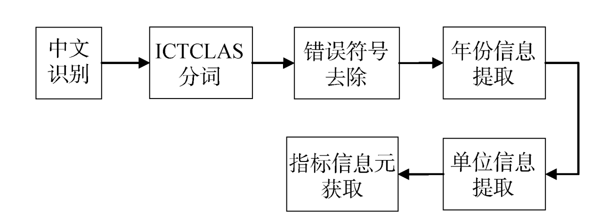


图 3 指标数据信息元提取流程

(1) 中文识别

指标文件中涵盖大量指标对应的英文翻译，这属于匹配无关信息，应予以去除，因此，需要对中文及英文字符块进行区分，提取中文字符块以进行后续操作。平台选取的中文编码方式为 Unicode，其中中文汉字的范围为 4E00 至 9FBF，可以通过字符的 ASCII 码的显示范围来确定是否为中文并进行中文的识别。而指标内的专有英文词汇如 GDP 等，属于指标信息元，应予以保留。

ChinaXiv合作期刊

chinaXiv:201711.02037v1

(2) 中文分词及错误符号的去除

指标描述的中文字符串内包含一些无用信息，如标点符号、助词、连词等。这些无用信息会干扰匹配结果，且影响匹配的效率。通过对指标的无用字符进行分析发现，利用词性对无关信息进行去除，能达到较好的效果。因此，本文利用分词系统对中文指标字符串进行分词，对分词的结果进行分析，根据词性去掉无关信息，从而得到表述指标的关键信息集合。

平台采用 ICTCLAS 分词系统，它可自定义词库，有良好的分词效果和强大的数据处理效率[13]，分析分词的结果，去除标点 w、助词 u、连词 c、介词 p 和量词 q 等这些对匹配无关的杂项。

(3) 年份及单位的提取

年份和单位是描述指标数据的重要参数，需要对其进行提取。表格中年份信息的提取较为简单，由于年份一般为数字，且通常出现于表格名称或行列标题，根据字符的数值范围即可判别是否为年份信息。在进行单位提取的时候，需要建立单位词表库，对年鉴表格中的单位信息进行识别，同时，对于表格中明确标明单位的单元格进行判别，扩充单位信息库。

对于表格中表名、行标题、列标题、单位这些描述指标的关键信息，综合上述步骤进行处理，即可去除无关中文字符杂项，得到描述单个指标的关键词集合，即指标信息元集合。特别需要注意，由于表名对指标的表述贡献明显弱于行列标题，因此本文按行标题、列标题、表名的顺序进行指标名称的提取，使此步骤的结果集按信息量权重排序，有利于提高指标的匹配效率。其中，指标年份和单位的信息，是进行指标数据的归一化合并所必需的要素。

3.3 匹配模式集合筛减

指标体系是海量数据的集合，对其进行筛减，可以减少匹配模式集合的数量，去除掉大量无关指标信息，保障了指标匹配的处理效率。指标体系存储在 MSSQL 数据库中，考虑到指标文本数量较大，此步骤采用数据库查询的方式进行，将指标的信息元作为关键词，对数据库进行 SQL 条件查询。在进行关键字搜索时常常使用 Like 运算符进行模糊查询，但是由于 Like 子句强制数据库系统线性扫描文本字段，降低了系统性能[14]。为了提高整个指标筛选过程的效率，本文利用数据库全文检索技术来实现数据的查询，采用 FullText 构造文本字符串中的单词索引，缩短了检索时间[14]。

由于该过程需要频繁操作数据库，本文采用存储过程实现该步骤。将指标信息元作为存储过程的参数传入，使大量 T-SQL 语句集合提前完成预编译和优化，直接返回筛选后的数据集合，从而提高执行速度。

具体的筛减过程如下：设需要匹配的文本为 A_1，对 A_1 进行信息元提取，得到集合 A{a_1，a_2，a_3，a_4，$a_5 \cdots a_n$}，通过集合 A 在数据库中进行条件查询，设定阈值为 10 000，将 a_1，a_2，a_3，a_4，$a_5 \cdots a_n$ 依次加入查询条件，写入循环，若查询不到指标数据，则去除无效关键词；若查询结果大于阈值，则继续循环；若查询结果小于阈值，则退出循环过程，筛选出相关指标集合 B，则集合 B 即为所有和 A_1 所在表格相关的指标名称集合，同时，为了避免后续重复查询，需要记录此次查询的条件集合 C，并以数据集的形式返回。

3.4 正向匹配处理

正向匹配处理是字符串匹配的过程，将指标信息元集合A{a_1，a_2，a_3，a_4，$a_5 \cdots a_n$}与筛减指标集合B{b_1，b_2，b_3，b_4，$b_5 \cdots b_m$}进行字符比对，设定阈值为100，查找出符合条件的最小字符集合B_1，并进行反向匹配。其主要处理思想与指标筛选过程类似，将a_1，a_2，a_3，a_4，$a_5 \cdots a_n$加入比对条件集合C时，若已存在集合C内，则继续循环；若不在，则加入查询条件集合C，由于信息元集合中有无效信息元出现，将条件集合C与指标集合B进行字符串匹配后，若无相关匹配指标返回，则需要去除该查询条件，若匹配指标数量小于阈值，则直接进行反向匹配处理。匹配流程主要实现伪代码如下所示：

```
for (i=0; i< A; i++) {
    char* sFilter= A[i];
    //若查询条件为空或查找条件已在查询集合内
    if (sFilter.IsEmpty()||sFilter in C){
        continue;
    }
    C.Add(sFilter);  //加入查询条件
    // WM 算法具体实现函数
    nRes=FindIndWM (C ,B);
    if (nRes ==0){
      //若查不到任何值，则该查询条件无效
      //去除该查询条件
    C.Remove(sFilter);
     continue;
     }else if(nRes>0&& nRes<100){
      //在阈值范围内
```

ChinaXiv合作期刊

chinaXiv:201711.02037v1

```
            //反向匹配
        }
    }
```

字符比对采用 WM 算法实现, 其实现思路见 2.3 节, 匹配过程的伪代码如下:

```
while(text<textend) {
    hashVal=hashBlock(text);//计算当前块的哈希值
    //查找块的坏字符移动表(SHIFT)得到下一个匹配
      开始位置
     shift_distance=SHIFT[hashval];
     //计算当前位置的哈希值
    if(shift_distance==0) { //当前块出现在某 pat 末
       shift_distance=1;
       p=HASH[hashval];
         //得到可能与当前块匹配的所有 pat 的集合的开始位置
       while(p)
       //检验子集中的 pat 是否匹配;
}
text+=shift_distance; //选择下一个可能的匹配入口
}
```

3.5 反向匹配处理

由于指标名称的相似性, 在实际匹配过程中发现仅仅进行正向匹配不足以解决实际的匹配问题。正向匹配一般会有多个指标与 A 完全匹配, 即便是集合 B_1 中仅仅只有唯一指标文本与 A 相匹配, 考虑到匹配数据对应的严谨性, 提高匹配准确性, 也要对匹配结果进行验证, 进行反向匹配。反向匹配是在正向匹配的基础之上, 对匹配结果的进一步筛选和验证。它与正向匹配相反, 是由指标集合 B_1 到录入指标字符串 A 的字符匹配过程。

指标进行正向匹配的结果, 往往有一个或者多个指标数据与需要匹配的指标相对应, 例如国民经济核算模块里面的指标:国际收支平衡表–经常项目–收益, 可以找到以下指标与之相对应:国际收支平衡表–经常项目–收益、国际收支平衡表–经常项目–职工报酬收益、国际收支平衡表–经常项目–投资收益等, 此时, 进行反向匹配, 可以去除掉错误匹配项国际收支平衡表–经常项目–职工报酬收益、国际收支平衡表–经常项目–投资收益, 如表 1 所示:

表 1 反向匹配流程示例

需要匹配指标	正向匹配的指标集	未匹配字符	结果
国际收支平衡表–经常项目–收益	国际收支平衡表–经常项目–收益	无	匹配
	国际收支平衡表–经常项目–职工报酬收益	职工报酬	不匹配
	国际收支平衡表–经常项目–投资收益	投资	不匹配

实现的思路如下:将正向匹配处理得到的数据集 B_1 进行信息元集合的提取, 并由它向需要匹配的指标名称 A 进行字符匹配对应, 若出现能找出唯一与之匹配的指标则匹配成功; 若找到多个匹配指标, 则匹配失败。如表 1 中, 职工报酬和投资不在需要匹配的指标国际收支平衡表–经常项目–收益中, 反向匹配结果为不匹配, 因此可以找到唯一匹配项, 则匹配成功。

4 实验分析与结论

平台的开发选择微软公司的 Visual C++ 2010, 采用 OLE/COM 实现对 Excel 表格的操作, 数据库服务器端选用的是 Microsoft SQL Server, 客户端同时支持 Microsoft SQL Server 和 Access。

为了验证双向匹配算法的应用效果, 对中国人口和就业统计年鉴 2010 的年鉴文件进行录入匹配测试工作, 操作界面如图 4 所示:

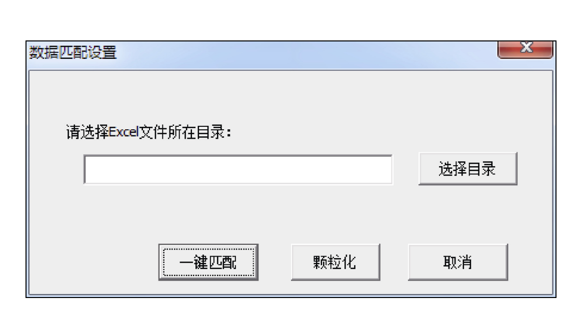


图 4 平台录入匹配操作界面

中国人口和就业统计年鉴 2010 共有 122 个年鉴文件, 对所有表格的指标数据进行匹配录入的操作, 其中, 成功 110 个文件, 失败 12 个, 平均单个表格的处理时间为 97 秒。失败 12 个文件中, 有 8 个是本身数据记录不连续(例如 2008 年 11 月 1 日至 2009 年 10 月 31 日一个时间段的数据)的特殊文件, 数据本身价值不大, 暂不需要采集录入, 还有另外 4 个文件表格指标读取有问题, 经分析该表格属于特殊表格, 需要调

ChinaXiv合作期刊

chinaXiv:201711.02037v1

整表格格式之后进行数据录入。

成功录入的 110 个文件中，含有指标共计 4 892 个，匹配成功的指标有 4 278 个。为了验证匹配结果的正确性，对结果进行错误检测，根据指标数据的特点，设计并实现了错误检测的程序，根据指标名称的特点，将疑似错误匹配到的指标挑选出来，如图 5 所示。通过程序自动进行错误检测，并进行人工批量审核，发现错误匹配到的指标有 678 个。错误指标可以通过平台查找关联指标的功能，进行匹配结果更正，如图 6 所示。对于已匹配的指标数据，需要审核之后批量入库；对于未匹配的指标数据，平台设计并实现了查找可能关联的指标功能，可以进行半自动人工关联。通过以上实验分析，双向匹配算法的匹配率达到 87.45%，正确率达到 86.14%。

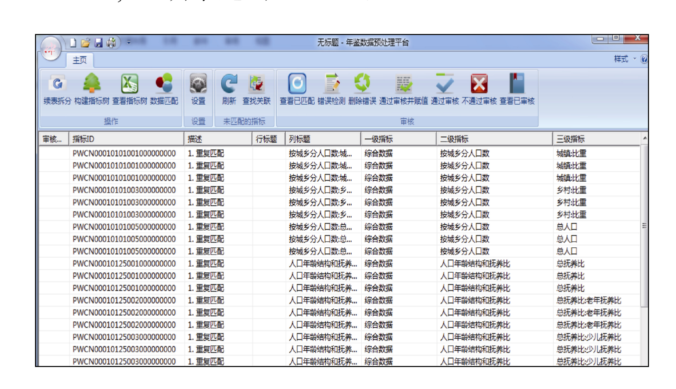


图 5　错误检测结果

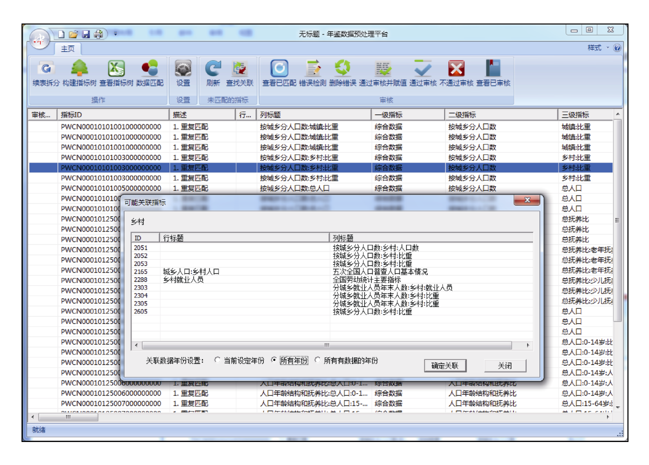


图 6　可能关联的指标列表

5　结　语

年鉴预处理平台，实现了指标数据的自动录入，使录入流程标准化，避免了人工录入可能出现的错误，大大节省了人力和时间成本，在年鉴数据的录入工作中起到重要的作用。而指标双向匹配处理，作为年鉴录入平台中的核心步骤之一，达到了较高的数据匹配准确率，实现了预期目标，节省了年鉴录入的时间和人力成本，保障了年鉴数据录入的准确性和安全性。但是为了达到较高的匹配率，在进行匹配处理时，在一定程度上牺牲了算法处理速度，在今后的工作中，将进一步分析指标自有规律，研究实现提高双向匹配算法的处理速度的方法，加强算法的实用性。

参考文献：

[1] 宋莉莉. 年鉴信息化的思考与探索[J] .兰台世界，2013(2): 11-12. (Song Lili. Consideration and Exploration on Informatics in Yearbook [J]. Lantai World, 2013(02): 11-12.)

[2] 樊胜. C/S 与 B/S 的结构比较及 Web 数据库的访问方式[J]. 情报科学，2001, 19(4): 443-445. (Fan Sheng. The Comparison Between C/S Structure and B/S Structure and the Ways to Access Web Database [J]. Information Science, 2001, 19(4): 443-445.)

[3] Alomari O, Othman Z. Bees Algorithm for Feature Selection in Network Anomaly Detection [J]. Journal of Applied Sciences Research, 2012(8): 1748-1756.

[4] 王春雨. 基于编辑距离的字符串模式匹配算法研究[D]. 秦皇岛：燕山大学，2015. (Wang Chunyu. The String Pattern Matching Algorithm Based on Edit Distance [D]. Qinhuangdao: Yanshan University, 2015.)

[5] Knuth D E, Morris Jr J H, Pratt V R. Fast Pattern Matching in String [J]. SIAM Journal on Computing, 1977, 6(2): 323-350.

[6] Boyer R S, Moore J S. A Fast String Searching Algorithm [J]. Communications of the ACM, 1977, 20(10): 762-772.

[7] Yao A C. The Complexity of Pattern Matching for a Random String [J]. SIAM Journal on Computing, 1979, 8(3): 368-387.

[8] Faro S, Lecroq T. The Exact Online String Matching Problem: A Review of the Most Recent Results [J]. ACM Computing Surveys (CSUR), 2013, 45(2): Article No.13.

[9] 侯淼. 并行串匹配算法研究[D]. 哈尔滨：哈尔滨工业大学，2014. (Hou Miao. Research of Parallel String Matching Algorithm [D]. Harbin: Harbin Institute of Technology, 2014.)

[10] Aho A V, Corasick M J. Efficient String Matching:An Aid to Bibliographic Search [J]. Communication of the ACM, 1975, 18(6): 333-340.

ChinaXiv合作期刊

[11] Wu S, Manber U. A Fast Algorithm for Multi-Pattern Searching[R]. Report TR-94-17. Tucson, AZ: Department of Computer Science, University of Arizona, 1994.
[12] 王一霈, 石春, 戴上静, 等. 一种改进的针对中文编码的 Wu-Manber 多模式匹配算法[J]. 小型微型计算机系统, 2015, 36(4): 779-781. (Wang Yipei, Shi Chun, Dai Shangjing, et al. An Improved Wu-Manber Multi-pattern Matching Algorithm for Chinese Encoding [J]. Journal of Chinese Computer Systems, 2015, 36(4): 778-781.)
[13] 张华平. ICTCLAS2011 接口文档[K]. 北京理工大学, 2011. (Zhang Huaping. ICTCLAS2011 API Document [K]. Beijing Institute of Technology, 2011.)
[14] 宋敏. 基于 SOA 图书馆数字资源整合平台关键技术的研究与实现[J]. 现代图书情报技术, 2009(9): 22-27. (Song Min. Research and Realization of Key Techniques of Library's Digital Resource Integration Platform Based on SOA [J]. New Technology of Library and Information Services, 2009(9): 22-27.)

作者贡献声明:

史礼婷, 钟永恒, 张骞: 提出研究思路, 设计研究方案;
胡思思, 李贞贞: 进行实验;
史礼婷, 胡思思, 张骞: 清洗和分析数据;
史礼婷: 论文起草;
史礼婷, 钟永恒: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: shilt@mail.whlib.ac.cn。
[1] 史礼婷. Yearbook2010.zip. 中国人口和就业统计年鉴 2010.
[2] 史礼婷. Inds2010.xlsx. 中国人口和就业指标分类体系.
[3] 史礼婷. yearbook_result2010.xlsx. 中国人口和就业统计年鉴 2010 指标匹配结果.

收稿日期: 2016-03-09
收修改稿日期: 2016-05-27

Using Bidirectional Pattern Matching Model to Pre-Process Yearbook Data

Shi Liting[1] Zhang Qian[2] Zhong Yongheng[1] Hu Sisi[1] Li Zhenzhen[1]
[1](Wuhan Library, Chinese Academy of Sciences, Wuhan 430071, China)
[2](The 9th Designing of China Aerospace Science Industry Corporation, Wuhan 430040, China)

Abstract: [**Objective**] We try to store the yearbook records as structured data, which will also be updated regularly. [**Context**] The yearbook data pre-process system is a C/S tool platform for collecting, auditing and uploading data. It was developed with VC++, and generated contents for the yearbook database. [**Methods**] We first modified the classic WM algorithm to build a new bidirectional pattern matching model. With the help of word segmentation technology, the new model could extract the metadata of original records. Then, we reduced the number of pattern sets with data storing procedure and bidirectional matched the records to ensure the effectiveness and efficiency of the system. [**Results**] The proposed algorithm achieved high level of matching rate and accuracy. [**Conclusions**] Bidirectional matching algorithm can meet the needs of the yearbook data entry, and improve the efficiency of the data preprocessing system.

Keywords: Bidirectional pattern matching The yearbook data WM algorithm

chinaXiv:201711.02037v1